AD

Award Number: DAMD17-02-1-0411

TITLE: Novel Membrane-Associated Targets for Diagnosis and Treatment of Breast Cancer

PRINCIPAL INVESTIGATOR: Brenton G. Mar
Carol Westbrook, Ph.D.

CONTRACTING ORGANIZATION: University of Illinois
Chicago, IL 60612-7205

REPORT DATE: May 2004

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

20041101 142

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
May 2004

3. REPORT TYPE AND DATES COVERED
Annual Summary (1 May 2003 - 30 Apr 2004)

4. TITLE AND SUBTITLE
Novel Membrane-Associated Targets for Diagnosis and Treatment of Breast Cancer

5. FUNDING NUMBERS
DAMD17-02-1-0411

6. AUTHOR(S)
Brenton G. Mar
Carol Westbrook, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Illinois
Chicago, IL 60612-7205

E-Mail: bar1@uic.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Original contains color plates: ALL DTIC reproductions will be in black and white

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

Proteins localized to the cell membrane or secreted show great promise as therapeutic targets and diagnostic markers because of their easy accessibility. However, determining protein localization by traditional methods is a difficult process. A "feature" of membrane-bound and secreted (MAS) proteins can be exploited to determine their membrane-bound status on a large scale. Because the mRNA transcripts of MAS proteins are translated in polysomes bound to the endoplasmic reticulum (ER), they can be separated from their heavier cytosolic counterparts by sucrose gradient centrifugation. At the end of year two, we have reproducibly separated the RNA of MCF7 cells into two fractions and hybridized them to Affymetrix microarrays. Using a training set of 881 MAS and cytoplasmic proteins, as annotated from SWIS-PROT, we show that genes with a membrane to cytoplasmic expression ratio over 1.08 are very likely to have MAS localization, with 97% specificity for those genes expressed above a threshold level. Applying these criteria to the remaining unknown and tentative localized genes on the microarray led to the identification of 810 predicted MAS genes. Combined with breast cancer expression and amplicon data, this could allow for the identification of potential novel membrane-bound and secreted drug targets and markers.

14. SUBJECT TERMS
Affymetrix GeneChip, microarray, membrane-bound polysomes, amplicon, biomarkers, gene discovery

15. NUMBER OF PAGES
11

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Novel Membrane-Associated Targets for Diagnosis and Treatment of Breast Cancer

Research Progress:

Task 1. **To analyze the membrane-bound and cytosolic fractions obtained from MCF7 cells on Affymetrix U133A GeneChips (Months 1-16): Complete**

a. Extract membrane-bound and cytosolic RNA from the breast cancer cell line MCF7 and evaluate the quality of the separation by real-time PCR (Month 1-12): Previously Reported Complete
b. Generate cRNA from each fraction and hybridize to Affymetrix Test Chip to ensure quality of each cRNA sample (Month 13-15): Complete
c. Hybridize each cRNA sample to the U133A set of GeneChips and analyze the data to generate a membrane/cytosolic (m/c) ratio for every expressed gene (Month 16): Complete

Instead of using the Affymetrix U95A microarrays as outlined in the original proposal, we upgraded to the Affymetrix U133A microarrays, which includes an updated probe set to reflect the changes in the Unigene 133 revision. 10μg of total RNA was labeled, hybridized to Affymetrix test arrays and human U133A microarrays, processed and scanned according to standard Affymetrix protocols. Each experiment was performed in triplicate. The resulting CEL data files were processed using the Bioconductor software suite (a set of libraries for R9). The Robust Multiarray Average (RMA) algorithm10-12 was used for normalization, background correction and expression value calculation. A membrane/cytosolic expression (MEM/CYT) ratio was determined for every gene.

Task 2. **Determine the predictive ability of this data set against both known membrane-bound and cytoplasmic proteins, and generate an annotated database of genes encoding proteins likely to be membrane-bound or secreted in MCF7 cells (Months 17-24): Complete**

a. Generate a reference set of membrane-bound and cytosolic proteins from online databases, such as Swiss-Prot (Month 17).

A reference set was created containing genes which encoded proteins that were annotated to have unambiguous membrane, secreted (MAS) or cytoplasmic localization in the Swiss-Prot database[13] and were represented on the Affymetrix U133A microarray. Each Affymetrix microarray element has a unique Affymetrix Probe ID that can be mapped to at least one SwissProt Accession number[14] . Each protein was then evaluated based on its Swiss-Prot "Cellular location" comment tag. Proteins were considered to have MAS localization if the tag contained one of the following identifiers: "Secreted", "Golgi", "Vesicular", "Membrane", "Lysosome", or "Peroxisome". Proteins were categorized as tentative if the tag also contained the terms, "Probable", "Possible", "Potential", and "By similarity", but considered unambiguous if not. Entries that contained "Nuclear", "Nucleus", and "Mitochondrial" were removed as there is some evidence that nuclear and mitochondrial proteins can be synthesized in either pathway and then imported into the organelle. This resulting set was then hand edited to remove entries containing multiple isoforms targeted for different subcellular compartments. Proteins are considered to have Cytoplasmic localization if the Swiss-Prot "Cellular location" tag contained "Cytoplasmic" or "Cytoplasm". Again, entries with "Probable", "Possible", "Potential", and "By similarity" were considered tentative, and entries containing "Nuclear", "Nucleus", and "Mitochondrial" were removed. This list was hand edited to remove any entries with multiple isoforms as well as entries that contained any references to membrane association or organelles

Task 2. Continued

b. Determine the predictive ability of the data obtained in Task 1 on the control lists. This includes experimenting with various normalization and statistical approaches. The MEM/CYT ratio threshold above which most genes will encode MAS proteins will be determined. The reference

set will be divided randomly into ten sets, and the ability to predict one set, based on threshold level calculated from the other nine will be evaluated for each set. The prediction will also be compared to computational methods when full sequence is available (Month 8-19).

At a given MEM/CYT expression ratio r, the probability of belonging to the membrane class p(m|r) is calculated by using Bayes' rule as shown in Equation 1.

$$p(m \mid r) = \frac{p(r \mid m)P_m}{p(r \mid m)P_m + p(r \mid c)P_c} \qquad (1)$$

where $p(r|m)$ and $p(r|c)$ is the proportion of MAS and cytoplasmic proteins, respectively, at MEM/CYT ratio r. The P_m and P_c factor corresponds to the prior probability of belonging to the MAS or cytoplasmic class. It is reasonable to expect that the prior probability of belonging to the cytoplasmic class would be much higher than the prior of belonging to the membrane class, however, since this calculation has never been done on breast cells (it is possible that this prior would change from tissue to tissue), a prior probability of 0.5 is the best compromise.

Sensitivity is defined as TP/(TP+FP), specificity is defined as TN/(FN+TN), and positive predictive value is defined as (TP/(TP+FN), where TP (true positives) are the number of MAS genes that are labeled correctly, FN (false negatives) are the number of MAS genes that are labeled incorrectly, TN (true negatives) are the number of cytoplasmic genes that are labeled correctly, and FP (false positives) are the number of cytoplasmic genes that are incorrectly labeled.

Because the distribution of MEM/CYT ratios for either class cannot be known a priori, it was necessary to train a classifier using a reference set of genes with known subcellular localization. Out of 22,283 elements in the Affymetrix U133A array, subcellular location annotation as described above was available for 8,912 elements. Unambiguous MAS annotation was found for 2,932 of these, while unambiguous Cytoplasmic annotation was found for 802 elements.

It is likely that only a subset of the elements on the U133A microarray will be expressed in MCF-7 cells at a level great enough for meaningful measurement. To determine that level, we evaluated our ability to distinguish known MAS genes from known Cytosolic genes in the reference set at varying total expression (E_T) levels, where $E_T \equiv$ MEM expression + CYT expression. A 10 fold cross- validation was performed, at increasing threshold levels of E_T, including only microarray elements with an E_T value greater than or equal to the threshold. Briefly, for each E_T level, the data was randomly partitioned into 10 groups, using 9 of these groups as a "training" set and the remaining group as a "testing" set. At each E_T level, the MEM/CYT ratio threshold was calculated (as described in above) for the training set for that E_T level. The positive predictive value, sensitivity, and specificity were calculated by examining the performance of predicting the testing set for that E_T level, and averages over the 10 groups were recorded. The performance of prediction for E_T thresholds ranging from 22,283 (100% of the microarray elements) to 1,106 (4.9% of microarray elements) was examined. The E_T level that corresponded to the highest sensitivity without a significant drop in positive predictive value or specificity was 633. At this level only 28% of probe sets with the highest E_T. are included, resulting in a final dataset of 6,239 probe sets that pass this threshold filtering. Of those, 531 have unambiguous MAS annotation and 350 are have unambiguous cytoplasmic annotation. Additionally, at this level, our 10 fold cross validation yields a 97.7% positive predictive value with 80% sensitivity and 97% specificity.

All of the 881 probe sets in the reference set with the E_T above the threshold of 633 were used to determine the MEM/CYT ratio that corresponds to the maximum posterior probability of belonging to the MAS class. The distribution of MEM/CYT ratios for genes with known localizations was examined (figure 1A) and it is interesting to note that the cytoplasmic genes show a discrete peak, while the MAS genes show a bimodal distribution with a smaller peak that associates with the cytoplasmic genes. An arbitrary cut-off MEM/CYT ratio of 1.08 was selected as giving the maximum posterior probability. Note that above this level, the majority of known cytoplasmic genes are excluded but a sizeable fraction of the MAS genes show a lower

MEM/CYT ratio. Thus, genes with a ratio below 1.08 cannot be designated with certainty as either MAS or cytoplasmic.

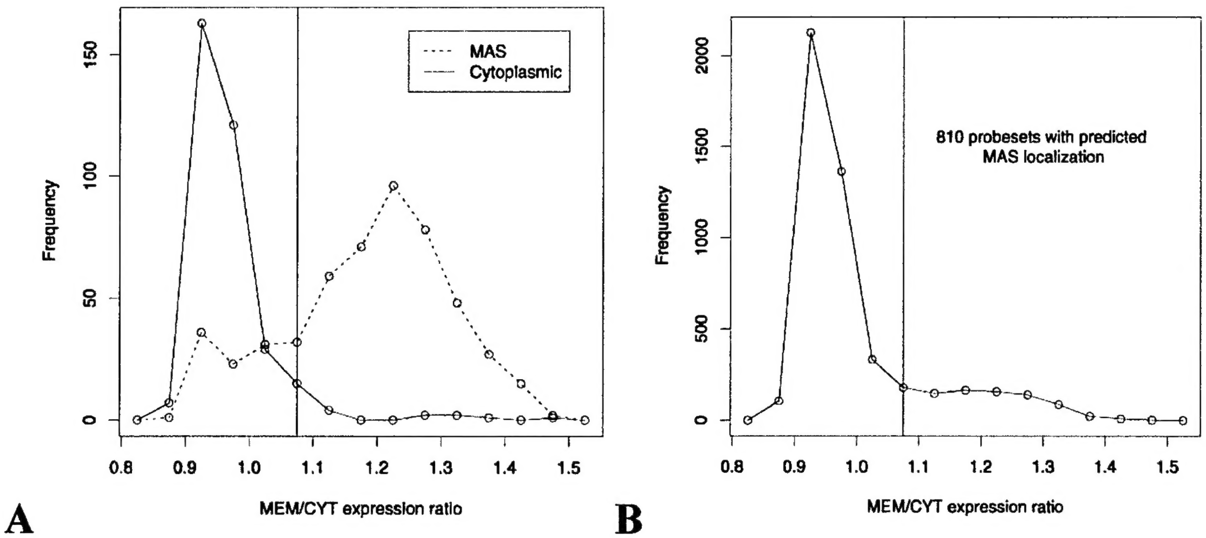

Figure 1A. Distribution of MEM/CYT ratios for all genes in the reference set expressing above the E_T. The vertical line indicates a MEM/CYT ratio of 1.08. B, **Distribution of MEM/CYT ratios for genes which are not in the reference set expressing above the E_T.** The vertical line indicates a MEM/CYT ratio of 1.08. At this threshold, 810 probesets were predicted to has MAS localization.

The distribution of MEM/CYT ratios for the remaining probesets (genes of unknown cellular localization) is plotted in Figure 1B. Of these, 810 probesets fall above the expression threshold and above the MEM/CYT ratio of 1.08. These 810 probesets are labeled as "Predicted MAS." Some of these genes are shown in Table 1. The remaining 4034 probesets found above the expression threshold and below the MEM/CYT ratio of 1.08 are labeled as "Indeterminant", because we expect a mixture of cytoplasmic and MAS genes in this range of MEM/CYT ratios. Figure 2 describes in graphical form of how these genes were selected the genechip. Of the "Predicted MAS" probesets, 343 were found to have a tentative subcellular annotation, but it did not meet the criteria previously established for the reference set. The remaining 467 probesets have no subcellular annotation. A similar percentage of "Indeterminant" probesets were found to have some tentative subcellular annotation (1535 out of 4034).

The SwissProt annotations were searched for terms that might indicate a tentative assignment to a cellular fraction (e.g. "Membrane by similarity" or "nuclear."). Over 70% (244 out of 343) of the predicted MAS probe sets with tentative annotations indicate a MAS subcellular location. Less than 5% (15 out of 343) of the predicted MAS probesets with tentative annotation are thought to be cytoplasmic. The remaining probe sets are thought to be localized to the nucleus, the endoplasmic reticulum, mitochondria and other intracellular organelles Biochemical process annotation was available for these 343 probe sets in Gene Ontology. Over half of these appear to be involved in metabolism, while a third are involved in cell growth. Almost 25% of the predicted MAS class are involved in cell communication. (While these annotations appear to comprise a greater number than the actual number of annotated probe sets, GO is organized in a way such that multiple annotations can correspond to a single probe set.)

In contrast, less than 15% (224 out of 1535) of the Indeterminate probe sets with tentative annotation indicate a MAS localization. Over 27% (425 out of 1535) of these are thought to be cytoplasmic. Interestingly, almost 50% of the indeterminant probesets contain nuclear annotation.

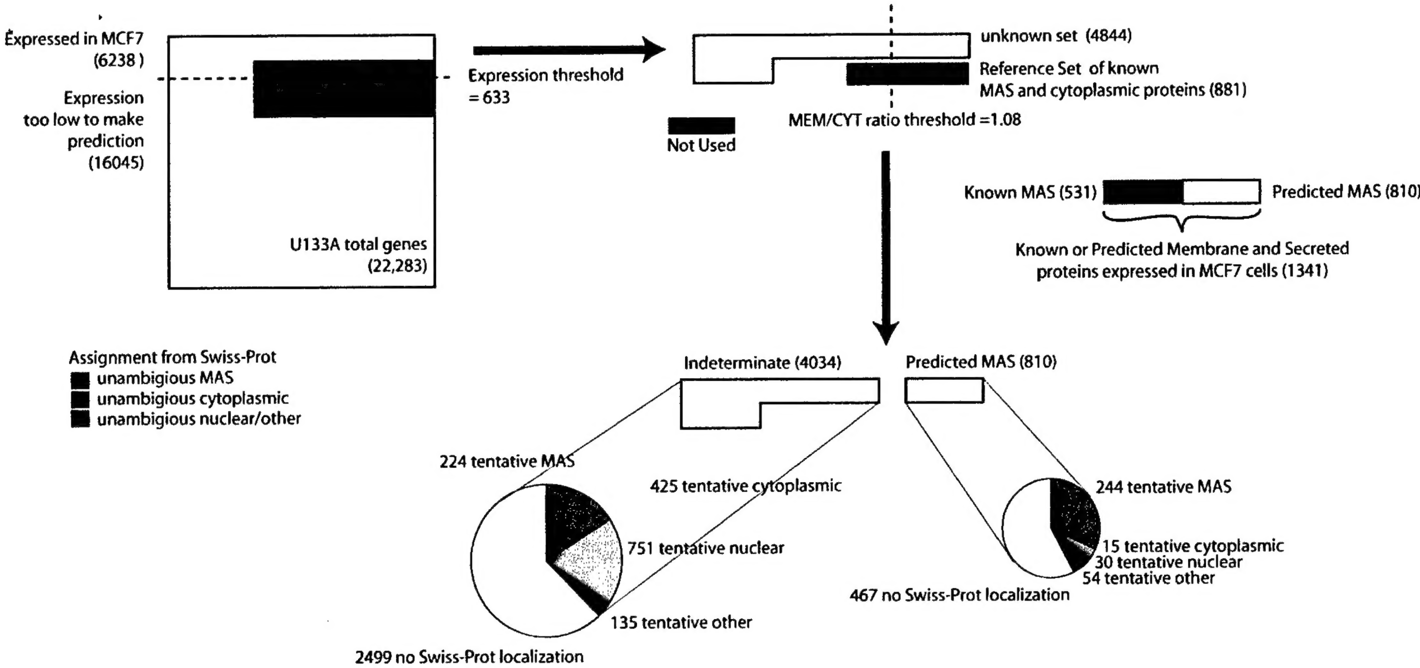

Figure 2. The relative expression of genes in the membrane and cytoplasmic RNA pools as measured by the Affymetrix 133A microarray is used to predict the membrane localization of genes expressed in MCF7 cells. An initial set of genes encoding proteins unambiguously labeled MAS or cytoplasmic in Swiss-Prot was identified, consisting of 2932 MAS and 802 cytoplasmic genes. Using this set, a MEM/CYT ratio threshold was determined, above which genes were predicted to be MAS, maximizing the specificity and sensitivity. To improve the prediction, we decided to exclude genes that were expressed at low levels, or not at all, in either RNA pool. The expression threshold was chosen empirically to maximize the sensitivity of the prediction without a significant drop in positive predictive value. Of the 22,283 elements on the microarray, 6238 or 28% passed the expression threshold, including 531 known MAS and 350 known cytoplasmic genes. These known expressed genes compose the Reference Set, from which the final MEM/CYT ratio threshold was determined. Applying the MEM/CYT ratio threshold to the unknown set of 4844 genes yielded 810 predicted MAS genes. Tentative localization from Swiss-Prot of these genes is shown. We have identified 1341 known and predicted MAS genes expressed in MCF7 cells.

Affymetrix ID	Accession #	Gene Name	Description	Localization (GO & SwissProt)	M/C Rat
212640_at	AF052159		Homo sapiens clone 24416 mRNA sequence	None	1.29
212248_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.26
212250_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.23
212251_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.21
201818_at	AF052162	FLJ12443	hypothetical protein FLJ12443	None	1.20
218686_s_at	Contig55188_RC	FLJ22341	hypothetical protein FLJ22341	None	1.11
219202_at	Contig55188_RC	FLJ22341	hypothetical protein FLJ22341	None	1.13
207170_s_at	NM_015416	HCCR1	cervical cancer 1 protooncogene	None	1.08
201037_at	D25328	PFKP	phosphofructokinase, platelet	None	1.11
208658_at	NM_004911	ERP70	protein disulfide isomerase related protein	ER	1.22
211048_s_at	NM_004911	ERP70	protein disulfide isomerase related protein	ER	1.26
210074_at	NM_001333	CTSL2	cathepsin L2	Lysosome	1.31
212290_at	AL050021		Homo sapiens mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.21
212295_s_at	AL050021		Homo sapiens mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.22
213094_at	AL080079	DKFZP 564D0462	hypothetical protein DKFZp564D0462	Membrane protein	1.34
219410_at	NM_018004	FLJ10134	hypothetical protein FLJ10134	Membrane protein	1.20
221675_s_at	NM_020244	LOC56994	cholinephosphotransferase 1	Membrane protein	1.35
210512_s_at	NM_003376	VEGF	vascular endothelial growth factor	Membrane protein	1.08
211527_x_at	NM_003376	VEGF	vascular endothelial growth factor	Membrane protein	1.10
203988_s_at	NM_004480	FUT8	fucosyltransferase 8 (alpha (1,6) fucosyltransferase)	Membrane protein (by similarity)	1.20

Table 1. A representative selection of the 810 genes or predicted to be MAS localized based on MEM/CYT ratios over 1.08

Task 2. ***Continued***

c. Additional data such as cytogenetic position, UniGene cluster number, and protein homology will be collected on each transcript. At this stage, we will generate an annotated database of genes encoding proteins likely to be membrane-bound or secreted in MCF7 cells. An annotated database of genes encoding cytosolic proteins will be generated as well (Month 20-24).

This task is currently in progress.

Task 3. **Identify genes encoding membrane-bound and secreted proteins that are known to be amplified, overexpressed, or differentially expressed in breast cancer. (Months 25-36): In progress.**

a. Use data from Task *2* to predict genes encoding membrane-bound and secreted proteins from amplicon data being generated in the mentor's lab from "genomic microarrays". Collect data from the literature supporting these candidates as potential drug targets and markers (Months 25-28).
b. Use data from Task 2 to predict genes encoding membrane-bound and secreted proteins from candidates identified in the literature. Collect data from the literature supporting these candidates as potential drug targets and markers (Months 29-32).
c. Develop data into an online public resource that breast cancer researchers can use to quickly screen their candidates for membrane-bound and secreted proteins (Months 33-36).

These tasks are currently in progress.

Key Accomplishments

Previous Reporting Period:

- Reliably and reproducibly separated RNA from the MCF7 cell line by discontinuous sucrose density gradient ultracentrifugation into membrane bound and cytoplasmic pools.
- Evaluated the quality and quantity of the RNA in each pool by spectrophotometry and gel analysis
- Tested two genes, GAPDH and JAM, by Realtime RT-PCR and found that they were appropriately enriched in each pool.

This Reporting Period:

- Hybridized RNA in membrane and cytoplasmic pools from sucrose gradient to Affymetrix U133A Genechips
- Constructed a reference set of genes whose proteins from SWIS-PROT that have been assigned unambiguous MAS (2,932) or cytoplasmic annotation (802).
- MEM/CYT ratios were calculated for all genes and the reference set was used to determine the ideal MEM/CYT ratio and expression level thresholds to best predict membrane and secreted protein localization. A gene with a MEM/CYT ratio above 1.08 is very likely to be a MAS protein, with a 97.7% positive prediction value, 80% sensitivity and 97% specificity in a ten-fold cross validation of the reference set.
- Applying these thresholds to the rest of the hybridization data, yields 810 genes that are "Predicted MAS" and will be looked at more closely with additional expression data.

Reportable Outcomes

Published abstract:

- Stitziel NO, Mar BG, Liang J, Westbrook CA. Membrane-associated and secreted proteins in the breast cancer transcriptosome. Proceedings of the AACR, Volume 45, Abstract 1655. (2004)

Manuscript in preparation:

- Membrane-Associated and Secreted Genes in Breast Cancer , Nathan O. Stitziel, Brenton G. Mar, Jie Liang, and Carol A. Westbrook

Conclusions

The work has progressed significantly since the last reporting period, with the membrane and cytoplasmic RNA pools successfully hybridized to Affymetrix 133A microarrays. An automated search of Swis-Prot yielded 2932 proteins with unambiguous MAS and 802 with unambiguous cytoplasmic annotation, which were represented on the microarrays. These were used in developing a reference set to determine the MEM/CYT expression ratio predictive of MAS localization using Bayesian statistics. Limiting the analysis to those genes that were well expressed in MCF7 cells, as opposed to all microarray elements, should aid the prediction. Instead of choosing an arbitrary n-fold expression over background cut-off, we were able to select a threshold empirically to maximize the positive predictive value of the MAS prediction without a large drop in sensitivity. This resulted in the inclusion of 6238 or 28% of the elements on the microarray, 4844 of which had ambiguous or unknown annotation in Swis-Prot, making the unknown set, and of which 881 had unambiguous MAS or cytoplasmic annotation, making the reference set. A ten-fold crossvalidation of the reference set yields a 97.7% positive predictive value with 80% sensitivity and 97% specificity of MAS prediction or those genes with MEM/CYT ratios above the 1.08 threshold, and it would not be unreasonable to expect that a prediction of the unknown would be of similar accuracy. However, there were a number of false negatives, reflected in the 80% sensitivity, because a portion of MAS genes had MEM/CYT ratios below the threshold. Because of this, genes falling below the threshold were labeled “indeterminate”, rather than given a localization. There a number of reasons for MAS genes with low thresholds, including alternate export mechanisms to the membrane and extracellular space, or polysome disassociation from the membrane during processing.

Applying the prediction to the 4811 genes in the unknown set yielded 810 predicted MAS genes. Together, with the 531 known MAS genes from Swis-Prot, make up a set of 1341 genes expressed in MCF7 breast cancer cells which encode proteins with known or highly probable membrane or secreted localization. In the next phase of the project, we will establish a publicly accessible database and identify candidates overexpressed or amplified in breast cancer that would make good candidates to study as markers or therapeutic targets.

References

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. 409(6822): p. 860-921.
2. Diehn M, Eisen MB, Botstein D, Brown PO. *Large-scale identification of secreted and membrane-associated gene products using DNA microarrays.* Nat Genet. 2000. 25(1):p. 58-62.
3. Mechler BM. *Isolation of messenger RNA from membrane-bound polysomes.* Methods Enzymol. 1987;152:p. 241-8.
4. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. 406(6797): p. 747-52.
5. Forozan, F., et al., *Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data.* Cancer Res, 2000. 60(16): p. 4519-25.
6. Hedenfalk, I., et al., *Gene-expression profiles in hereditary breast cancer.* N Engl J Med, 2001. 344(8): p. 539-48.
7. Pollack, J.R., et al., *Genome-wide analysis of DNA copy-number changes using cDNA microarrays.* Nat Genet, 1999. 23(1): p. 41-6.
8. Monni, O., et al., Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc Natl Acad Sci U S A, 2001. 98(10): p. 5711-6.
9. Ihaka, R. and Gentleman, R. (1996). "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics* 5, 299-314.
10. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan 22;19(2):185-93.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003 Apr;4(2):249-64.
12. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.
13. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003 Jan 1;31(1):365-70.
14. Downloadable at http://www.affymetrix.com/analysis/

APPENDICES

Abstract presented at the AACR meeting

Membrane-Associated and Secreted Proteins in the Breast Cancer Transcriptosome
N. Stitziel, B. Mar, J. Liang, C. Westbrook

Membrane-associated and secreted (MAS) proteins are one of the most important cellular components of a cancer, as they include transmembrane receptors and signaling molecules, adhesion molecules, and secreted autocrine peptides. They are also among the most useful targets for diagnosis and treatment, since their location makes them readily accessible to small molecule inhibitors as well as targeted antibodies. The objective of our study was to establish a database of MAS proteins which are present in a representative breast cancer cell line that could be a starting point for future target identification and validation. Because computational methods for finding MAS genes are limited, we used a biological approach, taking advantage of the fact that MAS proteins are preferentially translated in ribosomes associated with the endoplasmic reticulum, whereas cytosolic proteins are translated freely in the cytosol. Sucrose density centrifugation can be used to fractionate these ribosomes and prepare mRNA enriched for membrane-associated genes (MEM) and cytoplasmic genes (CYT). We used this approach to fractionate RNA from MCF-7 breast cancer cells, and the two fractions were labeled and hybridized in triplicate to Affymetrix U133A oligonucleotide microarrays. Expression ratios (MEM/CYT) were calculated for each of the 22,283 microarray elements by dividing the average signal from the MEM microarrays by the average signal from the CYT microarrays. We used the measured MEM/CYT ratio for 979 genes of known cellular location (772 membrane and 207 cytoplasmic genes) to calculate the probability of MAS class membership using Bayesian statistics, and established that the probabiliy was highest at or above a MEM/CYT ratio of 1.07. Applying this threshold to all of the data collected, we determined that 2,445 Affymetrix elements represent genes expressed in MCF-7 cells that are ≥94% likely to be membrane-associated or secreted, and thus comprise the Breast Cancer MAS (BCMAS) database. Studies are in progress to validate the components of the BCMAS database, and identify useful therapeutic targets including transmembrane kinases, receptors, and circulating peptides. To test the performance of our prediction threshold, 10 fold cross validation was performed. The training data was split randomly into 10 mutually exclusive test subsets. A threshold was calculated for each subset using the remaining training data and performance was calculated using the test subset. An average positive predictive value of 80% was achieved. Further work is underway to expand the training set in order to refine the threshold and increase the predictive value.